DOCUMENT RESUME

ED 282 424 FL 016 715

AUTHOR Kreeft, Henk; Sanders, Piet
TITLE Model Responses for Examinations with Open-Ended Questions.
PUB DATE May 83
NOTE 13p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706. Small print in figures may affect legibility.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Dutch; Evaluation Methods; Evaluators; Foreign Countries; *Interrater Reliability; *Language Tests; Models; *Native Language Instruction; *Reading Comprehension; *Reading Tests; Second Languages; Standardized Tests; Test Format; Testing Problems; *Test Items; Uncommonly Taught Languages
IDENTIFIERS *Netherlands

ABSTRACT

In the Dutch national examinations, reading comprehension tests are used for all languages. For the native language, reading comprehension is tested with reading passages and related questions to which the test-taker provides his own response, not choosing from a group of alternatives. One problem encountered in testing with these items is interrater reliability. In recent studies, raters have been provided with model responses, in more or less detailed forms, and the ratings on actual examinations have been analyzed. One major conclusion of the research has been that the influence of the raters' own standards was not controlled by the model responses provided. No matter how detailed the model responses were, the interrater reliability on those "open-ended" items was always rather low. Further research on how to control the influence of the raters' own standards is needed. Selecting and training raters for the national examinations might be appropriate. (MSE)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED282424

*At the Dutch national examinations reading comprehension tests are used for all languages. As regards the mother tongue, reading comprehension is tested by means of open-ended questions. Apart from the host of problems that arise at constructing the tests, such as:*
*- minding its context validity*
*- determining its degree of difficulty*
*- avoiding ambiguity of formulation*
*we also face the problem of rater reliability, which necessitates development of model responses and rating schemes. We will discuss some research we have done these last years on the rater reliability to be achieved with the help of different (i.e. more or less detailed) model responses. Experiments have been carried out and the results of the actual examinations have been analysed. One of the main conclusions is that the influence of the raters' own standards were not controlled by the model responses provided. No matter how detailed the modelresponse used, the rater reliability of the open-ended questions exams was always rather low. Further research on how to control the influence of the raters' own standards is needed. Selecting and training raters for the national examinations might be considered.*

In the Dutch national examinations two types of tests are used for the various subjects taught at school: the multiple-choice test and the test with open-ended questions. The reasons for this are based on political and traditional rather than on domainreferenced or testing criteria.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The multiple-choice exams are scored mechanically. Open-ended exams are scored by the students' own teacher and a second rater, a teacher at another school, who has been appointed by the authorities.
Compared to, for instance, England, this situation is rather unique: in the case of national examinations using open-ended questions the work of the candidates is never rated by their own teachers, but by officially appointed raters.

In recent years Cito has done some research on the problems connected with open-ended exams. In this introduction I should like to discuss three studies concerning open-ended reading tests in the mother tongue, in which interrater reliability plays a part.
From what follows it will become apparent that all three studies concentrate on test reliability and the role of raters in this respect. But the results as such cannot be compared because the research questions and analytical techniques differed considerably. A reason for presenting them together in this context is the fact that they give a fairly clear picture of the subject.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL016715



BEST COPY AVAILABLE

The first study concentrates on the aptitude of various types of rating-models for an experimental open-ended reading comprehension test.
The second study which is still in progress tries to develop an analytical rating model for a reading comprehension test comprising only one assignment: Give an approximately 400 word précis (summary) of a text of about 2500 words.
The third study focuses on test-reliability, including interrater reliability, as found in an actual open-ended reading comprehension exam.

A comparison will be made with similar studies on subjects such as physics, economics, music and geography. It is not possible to make a comparison with reading comprehension tests in foreign languages, such as French, German and English, as these tests are multiple-choice tests so that interrater reliability does not apply.

Question types

Before presenting the three individual studies, I think it is necessary to tell you something about the types of questions and rating-models we use for the various exams.

When considering the 'open'-ended questions that are used in the reading comprehension examinations we can see that the term 'open' perhaps is not correct for most of the questions. As a matter of fact many questions are rather closed. See for example questions 1, 2, 3 en 4. The answers we expect are often fully predictable. Many questions can be answered by giving a quotation from the text (question 1 and 2). Even when the question requires an answer to be formulated by the candidate this answer can usually be predicted by the testconstructor. So, when I speak of open-ended questions, I only mean that the candidate must write down the answer himself, unlike true-false questions, multiple-choice questions and matching questions. Real open-ended questions, which demand a non-predictable response from the candidate are hardly ever used.

In the appendix I have included some examples of questions as used in the examination of reading comprehension in the mother tongue. The text I used is an original English text from a second language examination. In the real examinations for the mothertongue the texts are much longer and the number of questions of course greater.

The first two questions are short-answer questions which are to be answered by quotation from the text or by indicating the passage in which the problem mentioned in the question is discussed.
The third question must be answered in the candidates' own words because it concerns an element of the text which is not explicit.
The fourth question is centered on the main idea of a passage in the text and should also be answered in the candidates' own words.

The fifth question asks for the candidate's judgement on a passage in the text (intra textual criticism) the question can be answered in different ways. The criterion here is not whether the answer is right or wrong, but the consistency between judgement and justification as well as the completeness of the answer.
The sixth question shows what a précis (summary) assignment is like. It is used for the testing of reading comprehension at pre-university level. Here too the original texts are much longer and so are the rating models (grids). In this assignment the following elements are tested:

a Has the student included the major points in the text and the relationships between them? (content factor)
b Is the students' phrasing correct? (the phrasing factor)
c Has the student structured his précis in a correct and well-ordered way? (the structure factor)

When testing with open-ended questions we have to face the problem of rater-reliability or rather unreliability. This means that our attention must also be directed to model responses and rating schemes. The position we take is that an open-ended question not finished after the question has been formulated. An open-ended question consists of the question and a detailed model response or rating-scheme.
A model response does not only consist of answers to be judged as right, but also of answers that are partly right, answers that are wrong and instructions about the deduction of scores.
A rating-scheme includes criteria according to which an assessment is made. And a description of the character of the answers going with each score.
Every type of question requires a specially adapted model for the assessment: multiple-choice questions and some types of short-answer questions and only need a correction key. Other short-answer questions and more extended questions require either a model response which lists all possible answers explicitly or a model response which gives indications as to the correctness of the answer in a more abstract form. For all those types of questions I have given examples in the appendix.

<u>The first study</u>: interrater reliability with a reading test in the mothertongue with open-ended questions at O-level.
In 1979 we did some research on rater reliability obtained by using various questions and model responses. Literature on rater reliability indicates that the rater reliability for open-ended questions is generally rather low. The accumulated evidence leads to three inevitable conclusions:

a different raters tend to assign different grades to the same paper
b a single rater tends to assign different grades to the same paper on different occasions
c the differences tend to increase as the essay questions permit greater freedom of response.

Most of the research however concerns the essay type

question. In the Dutch reading comprehension tests however we frequently use a question type in between short-answer questions and essay questions, which we call extended-answer questions. We wanted to know if the results of the research to be found in the literature concerned, also applied to these extended-answer questions.
Thus we constructed a test with different types of questions and different types of model responses:
short-answer questions - with explicit model responses
extended-answer questions - with explicit model responses
and extended-answer questions with more abstract model responses.

Of each reponse model we constructed two forms.
model response I with correct answers only and one or two answers which are partly correct. This model response was devised by the constructors of the test without knowing the actual answers of the candidates.
model response II which was more detailed and constructed after studying the actual answers of sixty candidates obtained by pretest.
We had two hypotheses:
1 the interrater reliability would be higher using the more detailed model responses than in the case of using the more simple model responses
2 the interrater reliability would be different for each type of question with both reponse models.

I shall not present the whole research plan here. I will restrict myself to saying that we worked with an experimental group of eight raters, thirty answers of candidates to each question. The experimental group judged the answers with model response I and three months later with model response II. A control group of 8 raters judged the answers with the more detailed reponse model II only. Interrater reliability was defined as a ratio of components of variance. The numerator of the ratio consists of the person-component. The denominator consists of the person-component the rater-component and the error-component.

These were our conclusions:
- There was no difference between the interrater reliability in using model response I or II
- Short-answer questions that only requested a quotation from the text had a rather high rater reliability .75 or more
- All questions that concern passages in the text of one sentence, a group of words or one word - irrespective of the type of question - had a rather high rater reliability, .75 or more
- All questions that concern more extended text passages had rather a low reliability of about .50
- It was interesting to learn that there was no difference in the interrater reliability between questions that had to be answered in the candidates' own words no matter if they were really open - (the criterion can't be whether the answer is right or wrong) - or if the answers were predictable
- Analysis of some specific questions showed that deduction rules (see appendix) phrased in general terms

had a negative effect on the interrater reliability and that deduction rules phrased in more explicit terms had a positive influence
- One of the most important conclusions was that the raters' own standards were not controlled by the model responses. The data showed raters who were rather consistently generous and raters who were rather consistently strict.

The fact that the results of the more detailed model response II, did not differ from the more simple model response I could be explained in many ways. Some questions for example could have been formulated in a better way. Some model responses became too extended and complex, and last but not least we did not know anything about the motivation of the raters. But in spite of all these problems I think it cannot be denied that making the model responses more detailed without any additional measures is not effective. The National examination could be the recommendation of tests with closed-questions (multiple choice, short-answer questions).
Other considerations however make this kind of decision inacceptable. We think of the validity of the test, the backwash effects of the test in the classroom and the problems of acceptation by the classroom teachers.
We suggest other possibilities:
- the questions, including extended-answer questions and essay questions, could be more structured, for example by adding response restrictions
- further research on the degree in which the raters own standards can be influenced. We would then know which instructions for raters would be more succesful than the instructions we already use.
- the training of raters could also be considered.

Similar studies were made with regard to physics and economics. In the case of physics some encouraging results were found. Assessment without the use of a model response on the basis of a short list of scores per question resulted in an interrater reliability of .70 and higher. The assessment on the basis of a more detailed model response resulted in an even higher interrater reliability.
In the case of economics however results similar to those in the Dutch reading comprehension test were found. There was no difference in assessment with the more detailed and the more simple reponse model.
In the case of economics half the questions had an interrater reliability of .70 or higher, one quater of the questions had an interrater reliability between .50 and .70 and another quarter even below .50.
These differences can only be explained by the differences between the types of questions the subjects give rise to.

In a second study of a model response the assessment of précis-writing was looked at. In an A-level précis-writing test a text of about 2500 words has to be summarized in 400 words. In the assessment of the results a schematic form of the text is used (see appendix, question 6).
For research purposes a model response was divised which consisted of a text-specific part: statements taken from

the text and their mutual relationships and two general parts in which phrasing and structure can be assessed on 5-point scales.
The rater reliability of 5 different raters was calculated on the basis of the mean intercorrelation of the raterscores lengthened by using the Spearman Brown formula.
The rater reliability obtained by using the text-specific analytical part of the model response was high: .80 which came close to the reliability in the case of global or jury assessment.
There was less agreement on the interpretation of the general criteria with regard to phrasing and structure. The rater reliability on these criteria was very low: between .20 and .40.
In a follow up study an attempt it will be tried to relate these criteria to the text specific model response.
It is true that the form of the exam question differs considerably from the exam questions in the studies referred to above. And the same holds good for the ways in which the coëfficients of the interrater reliability have been calculated. But the conclusion is the same: an acceptable rater reliability is out of the question when parts of the rating schemes ask for personal interpretation by the rater.
The experiences from the studies I have just described, were incorporated in the examinations and model responses. In a third study which I want to discuss now, the effect of the unreliability caused by the factor rater as one of the factors which influence the test-reliability of an open-ended test - the other factor being the candidates and questions - was studied.

In this study an O-level reading comprehension test of the mothertongue, that had been part of the 1980 national examination was looked at. The model response of the examination can be characterized as detailed. The work of 300 candidates was assessed by 15 pairs of independent raters, who each assessed the work of 20 candidates.

On the scores obtained in this way, an analysis of variance was carried out in which the candidates, the questions and the raters were looked at. The part of the observed variance which was caused by the factor rater appeared to be small, compared to the effect of the other factors, particularly the interaction between candidates and questions. The most important conclusion therefore is that the lack of test reliability is caused first of all by the other factors, rather than by disagreement of the raters. The reliability of the examaverages around .60; when the factor rater is incorporated the reliability averages .54. The judgement by one rater instead of two raters results in a test reliability of around .47.
Trying to enhance the reliability of these reading comprehension exams by resorting to model responses and multiple ratings is therefore hardly worthwile. A similar conclusion became even more manifest in a study of the assessment of music and geography (Doevendans, Post, Sanders, 1982). This study did not concentrate on rater reliability or test reliability but on comparison with a criterion. Scores on separate questions were not used in

the analysis: the emphasis was on total scores on the exams as a whole. Three assessment contexts were compared to a criterion, in this case: assessment by the classroomteacher and a second rater who knows the first rater's results, assessment by the classroomteacher only and an assessment in which an average was taken of the teacher's scores and the score of an independent rater. The average judgement of five independent raters was used as a criterion. The surprising result was that the total scores of the three assessment contexts hardly deviated from the criterion, even the judgement by just the classroomteacher. The results could be explained by the fact that the classroomteacher knew that a second rating would take place, the fact that detailed model responses were used and the fact that the exams mainly consisted of short-answer questions, which allowed only little variation between the answers.

When the results of the studies discussed above are considered, we arrive at a number of conclusions, which are hardly surprising:
the degree of interrater reliability depends on the questiontype which is determined by the subject.
But we also arrive at conclusions which demand further research or which could well be applied in our examinations.

- Parts of model responses specifically geared to the work which has to be assessed, have a positive effect on interrater reliability, while parts of model responses phrased in general terms have a contrary effect.

But this is small comfort to those who want to enhance the reliability of mother tongue examinations.

- We have seen that the lack of test reliability was not caused first of all by the disagreement of the raters, but that an extension of the number of questions will probably have a greater effect.

The fact that, in spite of model responses, raters stick to their own standards, that in every group of raters there are both mild and strict raters, may have hardly any effect on the examination as a whole, to the individual candidate it can be of the greatest importance. So further research on possibilities to influence these rater characteristics is needed.

I Television tapes, we believe, will give posterity a much better picture of 1974 than we have of, say, the 17th century, despite those indefatigable diarists. Yet the box does distort the image of the age in some ways.

II Size is the simplest illustration. When it all started the criticism heard most frequently was that the screen was too small and that the public, accustomed to the giants of the cinema screen, would never accept this world of dwarfs. In the early stages it was vain for insiders to argue that everything depended on distance from the screen and the angle of vision.

III Gradually the converted ceased to worry about the smallness but it remains a factor in producing curious effects.

IV The zoom lens and magnification produce other odd deceptions. Football pitches and, more particularly, cricket fields, tend to be foreshortened. People in the pavilion seem on top of the pitch. Creepy-crawlies become nightmare monsters, and the population revealed in a drop of pure drinking-water is a terrific boost for booze.

V These things are unimportant, it could be urged, and future generations will make allowances since the camera cannot lie. But can't it? It can't say anything at all if it is not there, a point that has struck me forcibly when watching the "World at War" series. I was impressed when the commentary honestly acknowledged that the scene as the two arms of the Russian pincers met round Stalingrad had been re-enacted but it made me wonder still more about the authenticity of other incidents.

VI When I discussed the question of screen distortion with a television engineer, he surprised me by saying that the spoken word was probably more authentic than the vision. This, he thought, was because producers were constantly seeking artistic effects and experimenting with angles that did not give a true perspective.

L. Marsland Gander in the Daily Telegraph

1 Quote a sentence from paragraphs I or II of which "Size" (line 4) is an illustration.

2 In paragraph III the writer mentions "curious effects".
In which other paragraph or paragraphs does the writer mention factors accounting for "curious effects"?

3 Explain that what the writer describes in paragraph IV is not a "lie" as meant in line 16.
Don't use more than 20 words.

4 What is the main idea of the passage?
Don't use more than 15 words.

5 a How do you judge the necessity of the second paragraph for the passage as a whole?

b State your reasons for your judgement.
Don't use more than 40 words.

6 Summarize this text in 100 words at most.

**Model response**

question 1

| | |
|---|---|
| *Yet the box does distort the image of the age in some ways* | 10 points |

no points:

any other sentence

question 2

| | |
|---|---|
| - *paragraph II* | 5 points |
| - *paragraph IV* | 5 points |

deduction:

deduct 4 points if more than two paragraphs are mentioned

question 3

the essence of a right answer should be:

*the author speaks of a lie when you don't know whether what is shown on television is reality or not* 10 points

for example:

partly correct:

*only a re-enacted incident, that is not presented as a simulation, is what the writer calls a lie* 5 points

*only a re-enacted incident is what the writer calls a lie* 2 points

deduction:

deduct 2 points if more than 25 words are used

no points:

- *on television many tricks are used*
- *the use of a zoomlens*
- *enlargements*
- *the size of the screen*

question 4

The essence of a correct answer should be:

*Television distorts reality* — 10 points

for example:

- *Television doesn't give a full authentic image of the incidents* — 10 points
- *Our opinion that television gives an authentic image of reality is not right* — 10 points
- *Television distorts the image of the age in some ways* — 10 points

partly correct:

- *Television doesn't give a much better picture of the incidents that occur nowadays than diarists did of the 17th century* — 7 points
- *With television you can never tell whether you are watching a re-enacted scene (of simulated incidents) or not* — 5 points
- *Television uses many tricks* — 3 points

deduction:

deduct 2 points if more than 20 words are used.

question 5

I a) *The second paragraph is needless*
b) A complete reason for this judgement should consist of:

1 *the main idea of the passage:*
2 *a comment on:*

*the size of the screen or people worried about the smallness of the screen in the prime years of television*

The relation between these two elements can be indicated as: when the main idea is 1. than 2. is unimportant.

for example:

*Because the author wants to tell us that nowadays television can show reality in a special perspective. In that case it is unimportant to know that in the prime years of television people worried about the smallness of the screen.* — 10 points

II a) *The second paragraph is very useful*
b) The essence of a complete reason for this judgement should be:

*The author first presents a line of arguments he doesn't endorse, in this line of arguments paragrah II has a function.* — 10 points

deduction:

deduct 2 points if more than 50 words are used.

Question 6

| line | Elements of information, structure and relationships* |
|---|---|
| 1 | Thesis: TV shows reality. |
| 2-3 | Antithesis: TV disturbs reality. |
| 4-10 | means: limited size of the screen<br>result: curious effects |
| 11-14 | means: enlargement by means of zoomlens<br>result: odd deceptions |
| 15-16 | assessment: irrelevant aspects to those who have become used to the fact that the camera does not lie. |
| 17-24 | distortion all the same<br><br>dramatic effects in documentaries<br>artistic effects (consciously achieved) |

* Only those relationships are dealt with which are explicit in the text.